Quantifying synergistic mutual information

Virgil Griffith[1,*] and Christof Koch[1,2]

[1]Computation and Neural Systems, Caltech, Pasadena, CA 91125
[2]Allen Institute for Brain Science, Seattle, WA 98103

Abstract

Quantifying cooperation among random variables in predicting a single target random variable is an important problem in many biological systems with 10s to 1000s of co-dependent variables. We review the prior literature of information theoretical measures of synergy and introduce a novel synergy measure, entitled *synergistic mutual information*, defined as the difference between the whole and the union of its parts. We apply all four measures against a suite of binary circuits to demonstrate that our measure alone quantifies the intuitive concept of synergy across all examples.

1 Introduction

Synergy is a fundamental concept in complex systems which that has received much attention in computational biology [1, 2]. Several papers [3–6] have proposed measures for quantifying synergy, but there remains no consensus which measure is most valid.

The concept of synergy spans many fields and theoretically could be applied to any non-subadditive function. But within the confines of Shannon information theory, synergy—or more formally, *synergistic information*—is a property of a set of n random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ cooperating to predict, that is reduce the uncertainty of, a single target random variable Y.

One clear application of synergistic information is in computational genetics. It is well understood that most phenotypic traits are influenced not only by single genes but by interactions among genes—for example, human eye-color is cooperatively specified by more than a dozen genes [7]. The magntitude of this "cooperative specification" is the synergistic information between the set of genes $\mathbf{X}$ and a phenotypic trait Y, here eye color. Another application is neuronal firings where potentially thousands of presynaptic neurons influence the firing rate of a single post-synaptic (target) neuron. Yet another application is discovering the "informationally synergistic modules" within a multi-scale complex system.

For pedagogical purposes all examples in the main text are *determinstic*, however, these methods equally apply to non-deterministic systems.

The prior literature [8, 9] has termed several distinct concepts as "synergy". This paper defines synergy as whole much the whole is greater than (the union of) its atomic elements (eq. (15)). The other notions of synergy are treated in our companion paper, "Quantifying the irreducibility of mutual information".

1.1 Notation

We use the following notation throughout. Let
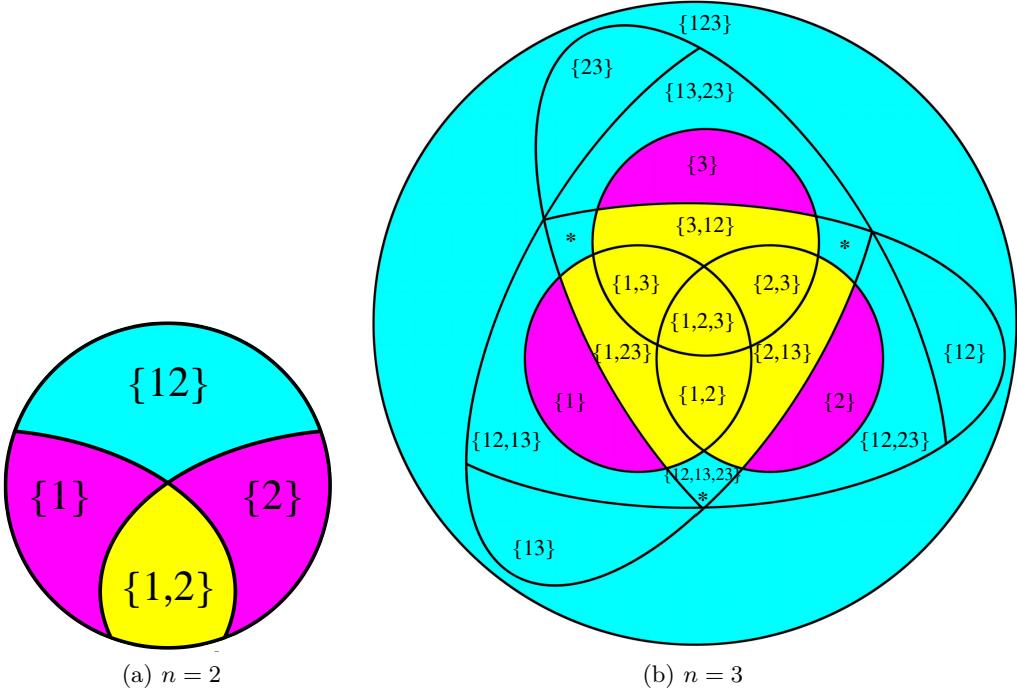
Figure 1: PI-diagrams for two and three predictors. Each PI-region represents nonnegative information about Y. A PI-region's color represents whether its information is redundant (yellow), unique (magenta), or synergistic (cyan). To preserve symmetry, the PI-region "$\{12, 13, 23\}$" is displayed as three separate regions each marked with a "*". All three *-regions should be treated as through they are a single region.

> n: The number of predictors $X_1, X_2, \ldots, X_n$. $n \geq 2$. In genetics, $X_1 \ldots X_n$ represent n distinct genes.
>
> $X_{1\ldots n}$: The *joint* random variable (coalition) of all n predictors $X_1 X_2 \ldots X_n$.
>
> X_i: The i'th predictor random variable (r.v.). $1 \leq i \leq n$.
>
> **X**: The *set* of all n predictors $\{X_1, X_2, \ldots, X_n\}$.
>
> Y: The *target r.v.* to be predicted. In genetics, Y represents a phenotypic trait (e.g. eye-color).
>
> y: A particular state of the target r.v. Y. In genetics, y is a particular state of the phenotype (e.g. eye-color = blue).

In this paper all random variables are discrete, all logarithms are $\log_2$, and all calculations are in *bits*. Entropy and mutual information are as defined by Shannon [10], $\mathrm{H}(X) \equiv \sum_{x \in X} \Pr(x) \log \frac{1}{\Pr(x)}$, and $\mathrm{I}(X : Y) \equiv \sum_{x,y} \Pr(x, y) \log \frac{\Pr(x,y)}{\Pr(x)\Pr(y)}$.

1.2 Understanding PI-diagrams

Partial information diagrams (PI-diagrams), introduced by Williams and Beer [6], extend Venn diagrams to properly represent synergy. Their framework has been invaluable to the evolution of our thinking on synergy.

A PI-diagram is composed of nonnegative *partial information regions* (PI-regions). Unlike the standard Venn entropy diagram in which the sum of all regions is the joint entropy $\mathrm{H}(X_{1\ldots n}, Y)$, in PI-diagrams the sum of all regions (i.e. the space of the PI-diagram) is the

mutual information $I(X_{1\ldots n}{:}Y)$. PI-diagrams are immensely helpful in understanding how the mutual information $I(X_{1\ldots n}{:}Y)$ is distributed across the coalitions and singletons of $\mathbf{X}$.[1]

How to read PI-diagrams. Each PI-region is uniquely identified by its "set notation" where each element is denoted solely by the predictors' indices. For example, in the PI-diagram for $n = 2$ (Figure 1a): $\{1\}$ is the information about Y only X_1 carries (likewise $\{2\}$ is the information only X_2 carries); $\{1, 2\}$ is the information about Y that X_1 as well as X_2 carries, while $\{12\}$ is the information about Y that is specified only by the coalition (joint random variable) $X_1 X_2$, while the entire disk corresponds to $I(X_1 X_2{:}Y)$.

The general structure of a PI-diagram becomes clearer after examining the PI-diagram for $n = 3$ (Figure 1b). All PI-regions from $n = 2$ are again present. Each predictor (X_1, X_2, X_3) can carry unique information (regions labeled $\{1\}$, $\{2\}$, $\{3\}$), carry information redundantly with another predictor ($\{1,2\}$, $\{1,3\}$, $\{2,3\}$), or specify information through a coalition with another predictor ($\{12\}$, $\{13\}$, $\{23\}$). New in $n = 3$ is information carried by all three predictors ($\{1,2,3\}$) as well as information specified through a three-way coalition ($\{123\}$). Intriguingly, for three predictors, information can be provided by a coalition as well as a singleton ($\{1,23\}$, $\{2,13\}$, $\{3,12\}$) or specified by multiple coalitions ($\{12,13\}$, $\{12,23\}$, $\{13,23\}$, $\{12,13,23\}$).

2 Information can be redundant, unique, or synergistic

Each PI-region represents an irreducible nonnegative slice of the mutual information $I(X_{1\ldots n}{:}Y)$ that is either:

1. **Redundant**. Information carried by a singleton predictor as well as available somewhere else. For $n = 2$: $\{1,2\}$. For $n = 3$: $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$, $\{1,23\}$, $\{2,13\}$, $\{3,12\}$.

2. **Unique**. Information carried by exactly one singleton predictor and is available no where else. For $n = 2$: $\{1\}$, $\{2\}$. For $n = 3$: $\{1\}$, $\{2\}$, $\{3\}$.

3. **Synergistic**. Any and all information in $I(X_{1\ldots n}{:}Y)$ that is not carried by a singleton predictor. $n = 2$: $\{12\}$. For $n = 3$: $\{12\}$, $\{13\}$, $\{23\}$, $\{123\}$, $\{12,13\}$, $\{12,23\}$, $\{13,23\}$, $\{12,13,23\}$.

Although a single PI-region is either redundant, unique, or synergistic, a single state of the target can have any combination of nonzero PI-regions. Therefore a single state of the target can convey redundant, unique, and synergistic information. This surprising fact is demonstrated in Figure 8.

2.1 Example Rdn: Redundant information

If X_1 and X_2 carry some identical[2] information (reduce the same uncertainty) about Y, then we say the set $\mathbf{X} = \{X_1, X_2\}$ has some *redundant information* about Y. Figure 2 illustrates a simple case of redundant information. Y has two equiprobable states: r and R (r/R for "redundant bit"). Examining X_1 or X_2 identically specifies one bit of Y, thus we say set $\mathbf{X} = \{X_1, X_2\}$ has one bit of redundant information about Y.

2.2 Example Unq: Unique information

X_i has *unique information* has about Y if and only if predictor X_i specifies information about Y that is not specified anywhere else (a singleton or coalition of the other $n - 1$

predictors). Figure 3 illustrates a simple case of unique information. Y has four equiprobable states: ab, aB, Ab, and AB. X_1 uniquely specifies bit a/A, and X_2 uniquely specifies bit b/B. If we had instead labeled the Y-states: 0, 1, 2, and 3, X_1 and X_2 would still have strictly unique information about Y. The state of X_1 would specify between {0,1} and {2,3}, and the state of X_2 would specify between {0,2} and {1,3}—together fully specifying the state of Y.

2.3 Example Xor: Synergistic information

A set of predictors $\mathbf{X} = \{X_1, \ldots, X_n\}$ has synergistic information about Y if and only if the whole $(X_{1\ldots n})$ specifies information about Y that is not specified by any singleton predictor. The canonical example of synergistic information is the Xor-gate (Figure 4). In this example, the whole $X_1 X_2$ fully specifies Y,

$$I(X_1 X_2 : Y) = H(Y) = 1 \text{ bit}, \tag{1}$$

but the singletons X_1 and X_2 specify *nothing* about Y,

$$I(X_1 : Y) = I(X_2 : Y) = 0 \text{ bits}. \tag{2}$$

With both X_1 and X_2 themselves having zero information about Y, we know that there can not be any redundant or unique information about Y—PI-regions $\{1\} = \{2\} = \{1, 2\} = 0$ bits. As the information between $X_1 X_2$ and Y must come from somewhere, by elimination we conclude that X_1 and X_2 synergistically specify Y.

3 Three examples elucidating properties of synergy

To aid the reader in developing intuition for any proper measure of synergy we illustrate some desired properties of synergistic information with pedagogical examples. All three examples derive from example Xor. Readers solely interested in the contrast with prior measures can skip to Section 4.

X_1 X_2	Y	
r r	r	$1/2$
R R	R	$1/2$

(a) $\Pr(x_1, x_2, y)$

(b) circuit diagram
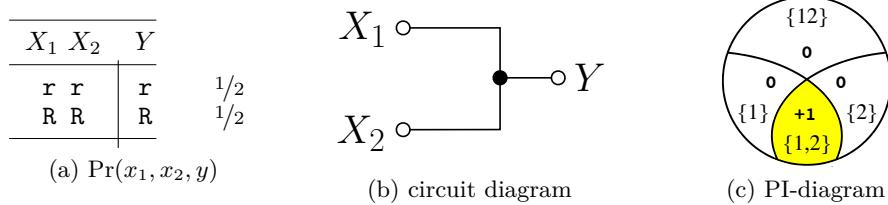
(c) PI-diagram

Figure 2: Example Rdn. Figure 2a shows the joint distribution of r.v.'s X_1, X_2, and Y, $\Pr(x_1, x_2, y)$, revealing that all three terms are fully correlated. Figure 2b represents the joint distribution as an electrical circuit. Figure 2c is the PI-diagram indicating that set $\{X_1, X_2\}$ has 1 bit of redundant information about Y. $I(X_1 X_2 : Y) = I(X_1 : Y) = I(X_2 : Y) = H(Y) = 1$ bit.

X_1 X_2	Y	
a b	ab	$1/4$
a B	aB	$1/4$
A b	Ab	$1/4$
A B	AB	$1/4$

(a) $\Pr(x_1, x_2, y)$

(b) circuit diagram

(c) PI-diagram
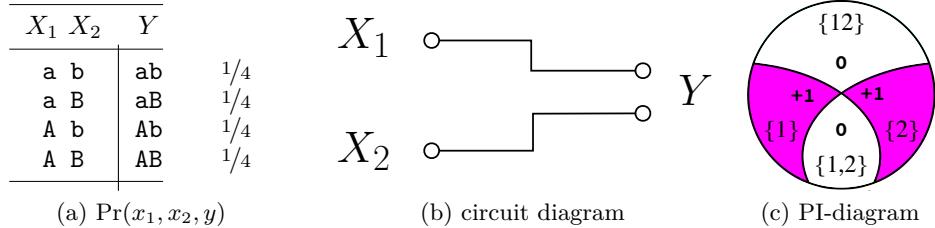
Figure 3: Example Unq. X_1 and X_2 each uniquely specify a single bit of Y. $I(X_1 X_2 : Y) = H(Y) = 2$ bits.

4

X_1	X_2	Y	
0	0	0	$1/4$
0	1	1	$1/4$
1	0	1	$1/4$
1	1	0	$1/4$

(a) $\Pr(x_1, x_2, y)$

X_1

XOR Y

X_2

(b) circuit diagram

{12} +1

0 0

{1} 0 {2}

{1,2}

(c) PI-diagram
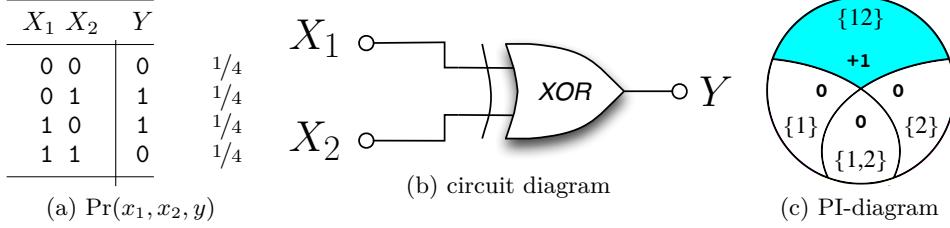
Figure 4: Example XOR. X_1 and X_2 synergistically specify Y. $I(X_1X_2:Y) = H(Y) = 1$ bit.

3.1 XorDuplicate: Duplicating a predictor does not change synergistic information

Example XORDUPLICATE (Figure 5) adds a third predictor, X_3, a copy of predictor X_1, to XOR. Whereas in XOR the target Y is specified only by coalition X_1X_2, duplicating predictor X_1 as X_3 makes the target equally specifiable by coalition X_3X_2.

Although now two different coalitions identically specify Y, mutual information is invariant to duplicates, e.g. $I(X_1X_2X_3:Y) = I(X_1X_2:Y)$ bit. Likewise for synergistic information to be likewise bounded between zero and the total mutual information $I(X_{1...n}:Y)$, synergistic information must similarly be invariant to duplicates, e.g. the synergistic information between set $\{X_1, X_2\}$ and Y must be the same as the synergistic information between $\{X_1, X_2, X_3\}$ and Y. This makes sense because if synergistic information is defined as the information in the whole beyond its parts, duplicating a part does not increase the net information provided by the parts. Altogether, we assert that *duplicating a predictor does not change the synergistic information*. Without the property that duplicating a predictor doesn't change synergistic information, the synergistic mutual information will not be bounded between 0 and $I(X_{1...n}:Y)$.

3.2 XorLoses: Adding a new predictor can decrease synergy

Example XORLOSES (Figure 6) adds a third predictor, X_3, to XOR and concretizes the distinction between synergy and "redundant synergy". In XORLOSES the target Y has one bit of uncertainty and just as in example XOR the coalition X_1X_2 fully specifies the target, $I(X_1X_2:Y) = H(Y) = 1$ bit. However, XORLOSES has *zero* intuitive synergy because the newly added singleton predictor, X_3, fully specifies Y by itself. This makes the synergy between X_1 and X_2 *completely redundant*—everything the coalition X_1X_2 specifies is now already specified by the singleton X_3.

4 Prior measures of synergy

4.1 I_{max} synergy: $\mathcal{S}_{max}(\mathbf{X}:Y)$

I_{max} synergy, denoted $\mathcal{S}_{max}$, derives from [6]. $\mathcal{S}_{max}$ defines synergy as the whole beyond the (state-dependent) *maximum* of its parts,

$$\mathcal{S}_{max}(\mathbf{X}:Y) \equiv I(X_{1...n}:Y) - I_{max}\left(\{X_1, \ldots, X_n\}:Y\right) \tag{3}$$

$$= I(X_{1...n}:Y) - \sum_{y \in Y} \Pr(Y=y) \max_i I(X_i:Y=y) , \tag{4}$$

where $I(X_i:Y=y)$ is [13]'s "specific-surprise",

$$I(X_i:Y=y) \equiv D_{KL}\left[\Pr(X_i|y) \middle\| \Pr(X_i)\right] \tag{5}$$

$$= \sum_{x_i \in X_i} \Pr(x_i|y) \log \frac{\Pr(x_i, y)}{\Pr(x_i)\Pr(y)} . \tag{6}$$

5

X_1 X_2 X_3	Y	
0 0 0	0	$1/4$
0 1 0	1	$1/4$
1 0 1	1	$1/4$
1 1 1	0	$1/4$

(a) $\Pr(x_1, x_2, x_3, y)$

X_1 XOR Y
X_2
X_3

(b) circuit diagram

(c) PI-diagram
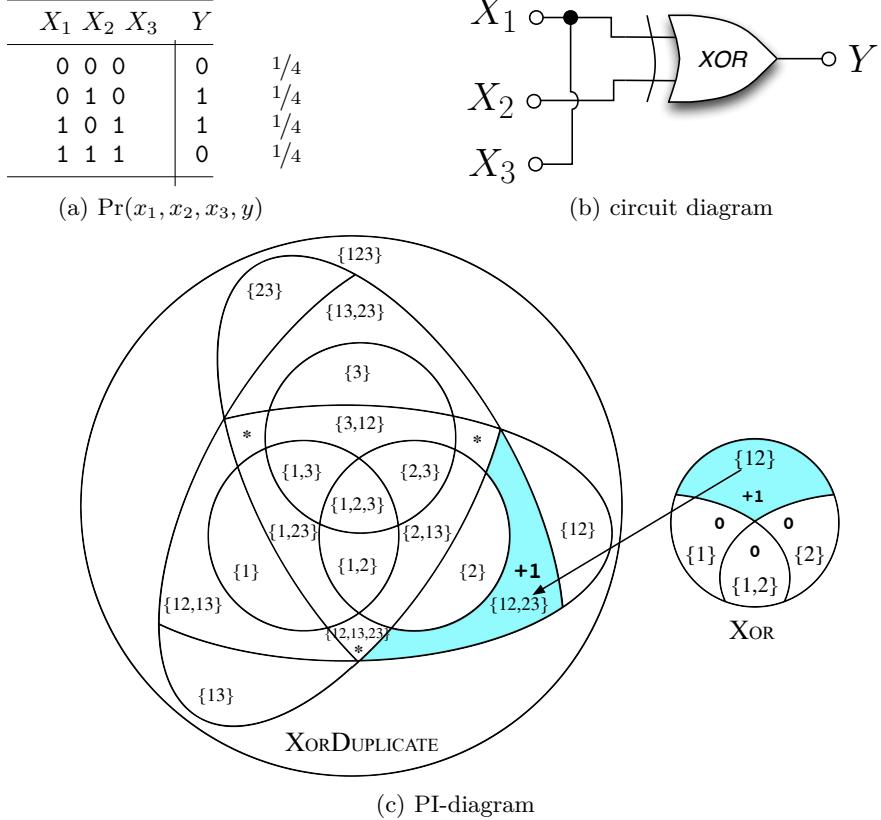
Figure 5: Example XORDUPLICATE shows that duplicating predictor X_1 as X_3 turns the single-coalition synergy $\{12\}$ into the multi-coalition synergy $\{12, 23\}$. After duplicating X_1, the coalition $X_3 X_2$ as well as coalition $X_1 X_2$ specifies Y. Synergistic information is unchanged from XOR, $\mathrm{I}(X_3 X_2 : Y) = \mathrm{I}(X_1 X_2 : Y) = \mathrm{H}(Y) = 1$ bit.

There are two major advantages of $\mathcal{S}_{\max}$ synergy. First, $\mathcal{S}_{\max}$ obeys the bounds of $0 \leq \mathcal{S}_{\max}(X_{1\ldots n} : Y) \leq \mathrm{I}(X_{1\ldots n} : Y)$. Second, $\mathcal{S}_{\max}$ is invariant to duplicate predictors. Despite these desired properties, $\mathcal{S}_{\max}$ miscategorizes merely unique information as synergistic whenever two or more predictors have unique information about the target. This can be seen in example UNQ (Figure 3). In example UNQ the wires in Figure 3b don't even touch, yet $\mathcal{S}_{\max}$ asserts there is one bit of synergy and one bit of redundancy—this is palpably strange.

The common defense of $\mathcal{S}_{\max}$ against example UNQ is to say one should "break up" Y into its components a/A and b/B and then compute $\mathcal{S}_{\max}$ for each component. Unfortunately this does not fully solve the problem because we often do not have the ability to "break up" Y. For instance, if the Y-states in UNQ were instead labeled as: 0, 1, 2, and 3, we wouldn't have the ability to break Y into its components.

A more abstract way to understand why $\mathcal{S}_{\max}$ would overestimate synergy—imagine a hypothetical example where there are exactly two bits of unique information for every state $y \in Y$ and no synergy or redundancy. $\mathcal{S}_{\max}$ would be the whole (both unique bits) minus the *maximum* over both predictors—which would be the max $[1, 1] = 1$ bit. The $\mathcal{S}_{\max}$ synergy would then be $2 - 1 = 1$ bit of synergy—even though by definition there was no synergy—but merely two bits of unique information.

Altogether, we conclude that $\mathcal{S}_{\max}$ *overestimates* the intuitive synergy by miscategorizing merely unique information as synergistic whenever two or more predictors have unique information about the target.

X_1 X_2 X_3	Y	
0 0 0	0	$1/4$
0 1 1	1	$1/4$
1 0 1	1	$1/4$
1 1 0	0	$1/4$

(a) $\Pr(x_1, x_2, x_3, y)$

(b) circuit diagram

(c) PI-diagram
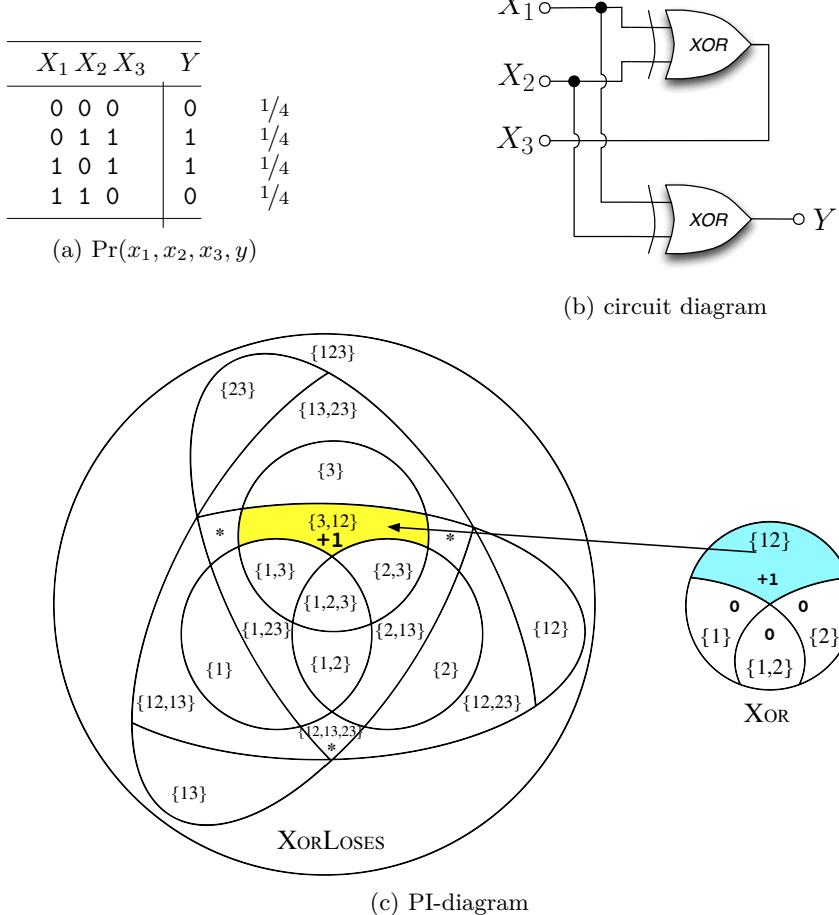
Figure 6: Example XorLoses. Target Y is fully specified by the coalition $X_1 X_2$ as well as by the singleton X_3. $\mathrm{I}(X_1 X_2 {:} Y) = \mathrm{I}(X_3 {:} Y) = \mathrm{H}(Y) = 1$ bit. Therefore the information synergistically specified by coalition $X_1 X_2$ is a redundant synergy.

4.2 WholeMinusSum synergy: $\mathrm{WMS}\,(\mathbf{X} : Y)$

The earliest known sightings of the bivarate case of WholeMinusSum synergy (WMS) is in [14, 15] and the general case in [16]. WholeMinusSum synergy is a signed measure where a positive value signifies synergy and a negative value signifies redundancy. WholeMinusSum synergy is defined by eq. (7) and interestingly reduces to eq. (10)—the difference of two *total correlations* (i.e. $\mathrm{TC}(X_1; \cdots ; X_n) = -\mathrm{H}(X_{1...n}) + \sum_{i=1}^{n} \mathrm{H}(X_i)$) [17].

$$\mathrm{WMS}\,(\mathbf{X} : Y) \equiv \mathrm{I}(X_{1...n} {:} Y) - \sum_{i=1}^{n} \mathrm{I}(X_i {:} Y) \tag{7}$$

$$= \mathrm{H}(X_{1...n}) - \mathrm{H}(X_{1...n}|Y) - \sum_{i=1}^{n} \mathrm{H}(X_i) + \sum_{i=1}^{n} \mathrm{H}(X_i|Y) \tag{8}$$

$$= \mathrm{TC}\,(X_1; \cdots ; X_n|Y) - \mathrm{D_{KL}}\left[\Pr(X_{1...n}) \middle\| \prod_{i=1}^{n} \Pr(X_i)\right] \tag{9}$$

$$= \mathrm{TC}\,(X_1; \cdots ; X_n|Y) - \mathrm{TC}\,(X_1; \cdots ; X_n) \tag{10}$$

Writing eq. ([7]) for $n = 2$ as a PI-diagram (Figure [7a]) reveals that for $n = 2$ WMS is the synergy between X_1 and X_2 *minus* their redundancy. Thus, if there were an equal magnitude of synergy and redundancy between X_1 and X_2 (as in RDNXOR, Figure [8]), WholeMinusSum synergy would be *zero*—leading one to *erroneously* conclude there is no synergy or redundancy present.[3] WholeMinusSum's PI-diagram for $n = 3$ (Figure [7b]) reveals that for $n > 2$, WMS $(\mathbf{X} : Y)$ becomes synergy minus the redundancy *counted multiple times* (the example PARITYRDNRDN in Appendix [A] demonstrates this).

Thus WholeMinusSum *underestimates* the intuitive synergy for all n with the potential gap increasing with n. Equivalently, we say that WholeMinusSum synergy is a *lowerbound* on the intuitive synergy with the bound becoming looser with larger n. For example, for $n = 2$ (Figure [7a]) WholeMinusSum double-subtracts PI-region {1,2}, but for $n = 3$ (Figure [7b]) WholeMinusSum double-subtracts PI-regions {1,2}, {1,3}, {2,3} and triple-subtracts PI-region {1,2,3}.

(a) WMS $\left(\{X_1, X_2\} : Y\right)$ (b) WMS $\left(\{X_1, X_2, X_3\} : Y\right)$
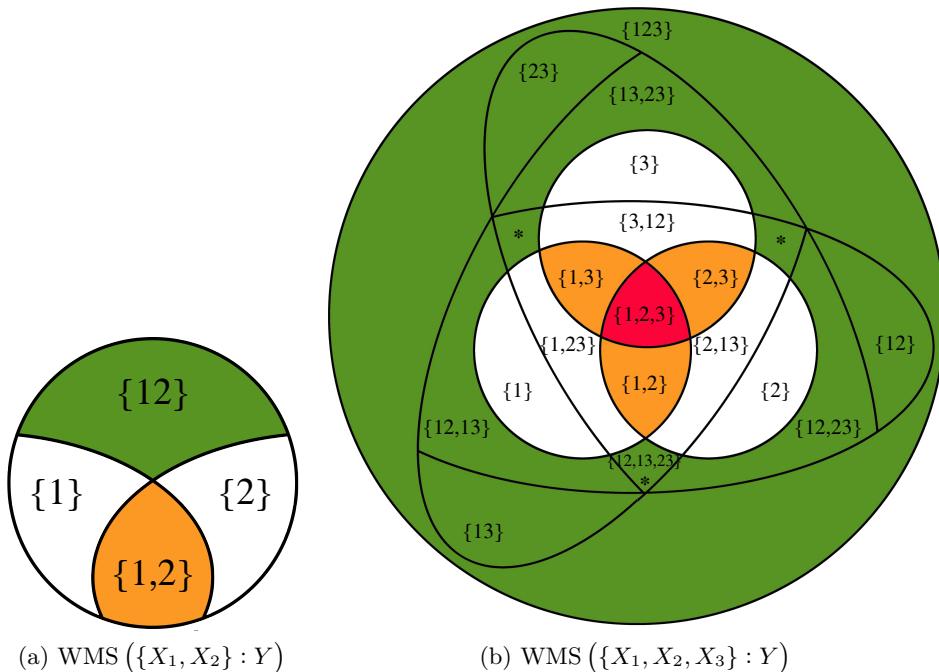
Figure 7: PI-diagrams representing WholeMinusSum synergy for $n = 2$ (left) and $n = 3$ (right). For this diagram the colors merely denote the added and subtracted PI-regions. WMS $(\mathbf{X} : Y)$ is the green PI-region(s), minus the orange PI-region(s), minus two times any red PI-region.

A concrete example demonstrating WholeMinusSum's "synergy minus redundancy" behavior is example RDNXOR (Figure [8]) which overlays examples RDN and XOR to form a single system. The target Y has two bits of uncertainty or entropy, i.e. $\mathrm{H}(Y) = 2$. Like RDN, either X_1 or X_2 identically specifies the letter of Y (r/R), making one bit of redundant information. Like XOR, only the coalition $X_1 X_2$ specifies the digit of Y (0/1), making one bit of synergistic information. Together this makes one bit of redundancy and one bit of synergy.

Note that in RDNXOR every state $y \in Y$ conveys one bit of redundant information and one bit of synergistic information, e.g. for the state $y = \mathtt{r0}$ the letter "$\mathtt{r}$" is specified redundantly and the digit "$\mathtt{0}$" is specified synergistically. Example RDNUNQXOR (Appendix [A]) extends

---

[3]This is different from [3]'s point that a mish-mash of synergy and redundancy across different states of $y \in Y$ can average to zero. Figure [8] evaluates to zero for *every state $y \in Y$*.

RDNXOR to demonstrate redundant, unique, and synergistic information for every state $y \in Y$.


X_1 X_2	Y	
r0 r0	r0	1/8
r0 r1	r1	1/8
r1 r0	r1	1/8
r1 r1	r0	1/8
R0 R0	R0	1/8
R0 R1	R1	1/8
R1 R0	R1	1/8
R1 R1	R0	1/8

(a) $\Pr(x_1, x_2, y)$

(b) circuit diagram

(c) PI-diagram
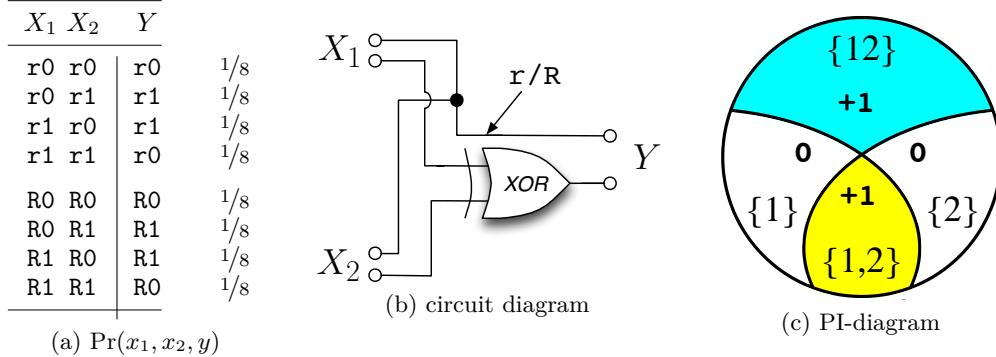
Figure 8: Example RDNXOR has one bit of redundancy and one bit of synergy. For this example, $\mathrm{WMS}(\mathbf{X} : Y) = 0$ bits.

4.3 Correlational importance: $\Delta\,\mathrm{I}\,(\mathbf{X}; Y)$

Correlational importance, denoted $\Delta\,\mathrm{I}$, comes from [5, 18–21]. Correlational importance quantifies the "informational importance of conditional dependence" or the "information lost when ignoring conditional dependence" among the predictors decoding target Y. As conditional dependence is necessary for synergy, $\Delta\,\mathrm{I}$ seems related to our intuitive conception of synergy. $\Delta\,\mathrm{I}$ is defined as,

$$\Delta\,\mathrm{I}\,(\mathbf{X}; Y) \;\equiv\; \mathrm{D_{KL}}\Big[\Pr\big(Y|X_{1\ldots n}\big)\Big\|\Pr_{\mathrm{ind}}\big(Y|\mathbf{X}\big)\Big] \tag{11}$$

$$= \sum_{y, \mathbf{x} \in Y, \mathbf{X}} \Pr(y, x_{1\ldots n}) \log \frac{\Pr\big(y|x_{1\ldots n}\big)}{\Pr_{\mathrm{ind}}(y|\mathbf{x})} \;, \tag{12}$$

where $\Pr_{\mathrm{ind}}\big(y|\mathbf{x}\big) \equiv= \frac{\Pr(y)\prod_{i=1}^{n}\Pr\big(x_i|y\big)}{\sum_{y'}\Pr(y')\prod_{i=1}^{n}\Pr\big(x_i|y'\big)}$. After some algebra[4] eq. (12) becomes,

$$\Delta\,\mathrm{I}\,(\mathbf{X}; Y) = \mathrm{TC}\,(X_1; \cdots ; X_n|Y) - \mathrm{D_{KL}}\left[\Pr(X_{1\ldots n})\left\|\sum_{y}\Pr(y)\prod_{i=1}^{n}\Pr\big(X_i|y\big)\right.\right] , \tag{13}$$

which strikingly resembles WholeMinusSum eq. (9) reproduced below,

$$\mathrm{WMS}\,(\mathbf{X} : Y) = \mathrm{TC}\,(X_1; \cdots ; X_n|Y) - \mathrm{D_{KL}}\left[\Pr(X_{1\ldots n})\left\|\prod_{i=1}^{n}\Pr(X_i)\right.\right] .$$

Eqs. (9) and (13) have the same upperbound of $\mathrm{TC}\,(X_1; \cdots ; X_n|Y)$ and furthermore are algebraically identical up to the righthand-side of the KL-divergence. Such uncanny similarities has led some to think that $\Delta\,\mathrm{I}$ quantifies some kind of synergistic information; indeed, there has been heated debate [3, 21] contrasting WMS and $\Delta\,\mathrm{I}$.

$\Delta\,\mathrm{I}$ is conceptually innovative and moreover agrees with our intuition for almost all of our examples. Yet further examples reveal that $\Delta\,\mathrm{I}$ measures something ever-so-subtly different from intuitive synergistic information.

---

[4]See Appendix B for the algebraic steps between eqs. (12) and (13).

9

The first example is [3]'s Figure 4 where $\Delta\,\mathrm{I}$ exceeds[5] the mutual information $\mathrm{I}(X_{1\dots n}\!:\!Y)$ with $\Delta\,\mathrm{I}\,(\mathbf{X};Y) = 0.0145$ and $\mathrm{I}(X_{1\dots n}\!:\!Y) = 0.0140$. This fact alone prevents interpreting $\Delta\,\mathrm{I}$ as a loss of mutual information from $\mathrm{I}(X_{1\dots n}\!:\!Y)$. Although $\Delta\,\mathrm{I}$ can not be a loss of mutual information, it could still be a loss of some alternative information (like Wyner's common information [22, 23]).

Instead, could $\Delta\,\mathrm{I}$ upperbound synergy? We turn to example AND (Figure 9). Example AND has $n = 2$ independent predictors and target Y is the AND of X_1 and X_2. Although AND's PI-region decomposition is subtler than XOR, we can still intuit its decomposition by a fortunate special case.

For X_1 and X_2 to redundantly specify Y, X_1 and X_2 themselves must have some information about each other.[6] However, because X_1 and X_2 are independent, $\mathrm{I}(X_1\!:\!X_2) = 0$ bits, there must be *zero* redundant information—meaning PI-region $\{1,2\} = 0$ bits.[7] With zero redundancy, the unique information PI-regions are simply the mutual information between the singletons and the target, $\{1\} = \mathrm{I}(X_1\!:\!Y) = 0.311$ bits and $\{2\} = \mathrm{I}(X_2\!:\!Y) = 0.311$ bits—these are computed using the uniform distribution per Figure 9a. From there, the synergy (PI-region $\{12\}$) is simply the whole, $\mathrm{I}(X_1 X_2\!:\!Y)$, minus the unique PI-regions ($\{1\}$ and $\{2\}$) and redundant PI-region ($\{1,2\}$) for $0.811 - 0.311 - 0.311 = 0.189$ bits of synergy.


(a) $\mathrm{Pr}(x_1, x_2, y)$

(b) circuit diagram

(c) PI-diagram
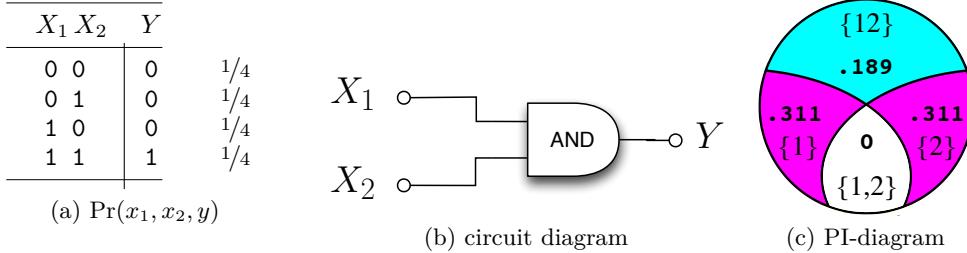
Figure 9: Example AND. X_1 and X_2 each have 0.311 bits of unique information. Additionally, X_1 and X_2 synergistically specify 0.189 bits, and redundantly specify zero bits. $\mathrm{I}(X_1 X_2\!:\!Y) = \mathrm{H}(Y) = 0.811$ bits.

In example AND the WMS synergy—the *lowerbound* on the intuitive synergy—is ≈ 0.189 bits, yet $\Delta\,\mathrm{I}\,(\mathbf{X};Y) = 0.104$ bits, and we conclude that $\Delta\,\mathrm{I}$ does not upperbound synergy.

Finally, in the face of duplicate predictors $\Delta\,\mathrm{I}$ often *decreases*. From example AND to ANDDUPLICATE (Section 4.4, Figure 10) $\Delta\,\mathrm{I}$ drops 63% to 0.038 bits.

Taking all three examples together, we conclude $\Delta\,\mathrm{I}$ measures something fundamentally different from synergistic information.

4.4 AndDuplicate: One example to rule them all

Our final example, ANDDUPLICATE (Figure 10), reveals undesirable behavior in all three prior measures. Example ANDDUPLICATE adds a duplicate predictor to example AND to show how each synergy measure responds to a duplicate predictor in a less pristine example than XOR. Before in XORDUPLICATE, we saw that when duplicating predictor X_1, the synergistic information *was unchanged*. But unlike XOR, in example AND both X_1 and X_2 have unique information—what happens to those two unique informations when duplicating a predictor? Most importantly, would either reduce synergy in the spirit of XORLOSES? Taking each one at a time:

---

[5]As $\Delta\,\mathrm{I}\,(\mathbf{X};Y)$ is often normalized by $\mathrm{I}(X_{1\dots n}\!:\!Y)$, it's concerning that $\Delta\,\mathrm{I}\,(\mathbf{X};Y)$ can *exceed* $\mathrm{I}(X_{1\dots n}\!:\!Y)$.

[6]A way to conceptualize this is that for two predictors to have redundant information about a target, the two predictors themselves must have some overlapping/redundant entropy, for two independent predictors this is $\mathrm{H}(X_1) + \mathrm{H}(X_2) - \mathrm{H}(X_1 X_2) = 0$ overlapping bits.

[7]Our assumption that positive redundant information between X_1 and X_2 requires positive $\mathrm{I}(X_1\!:\!X_2)$ is disputed by [24]. A forthcoming publication [25] gives credence to our assumption.

- Predictor X_2 is unaltered from example AND. Thus X_2's unique information stays the same. AND's $\{2\} \to$ ANDDUPLICATE's $\{2\}$.

- Predictor X_3 is identical to X_1. Thus all of X_1's unique information in AND becomes redundant information between predictors X_1 and X_3. AND's $\{1\} \to$ ANDDUPLICATE's $\{1,3\}$. When duplicating a predictor, the predictor's unique information becomes redundant information.

- In AND there is synergy between X_1 and X_2, and this synergy is still present in ANDDUPLICATE. Just as in XORDUPLICATE, the only difference is that now an identical synergy also exists between X_3 and X_2. Thus AND's $\{12\} \to$ ANDDUPLICATE's $\{12,23\}$.

- Predictor X_3 is identical to X_1. Therefore any information in AND that is specified by both X_1 and X_2 would now be specified by X_1, X_2, and X_3. Thus AND's $\{1,2\} \to$ ANDDUPLICATE's $\{1,2,3\}$.

Inspecting the finished PI-diagram in Figure 10c we see that duplicating a predictor leaves the intuitive synergistic information unchanged. To what conclusions do the three measures of synergy in the literature come?

WholeMinusSum synergy arrives at 0.189 bits of synergy for AND, but $0.189 - 0.311 = -0.123$ bits for ANDDUPLICATE. This again shows that WholeMinusSum subtracts redundancy from synergy, and is not invariant to duplicate predictors.

Correlational importance, $\Delta\,\mathrm{I}$, arrives at 0.104 bits of synergy for AND, but 0.038 bits for ANDDUPLICATE. $\Delta\,\mathrm{I}$ is not invariant to duplicate predictors.

$\mathcal{S}_{\max}$ arrives at 0.5 bits of synergy for both AND and ANDDUPLICATE. This is expected because $\mathcal{S}_{\max}$ is provably invariant to duplicate predictors—the only problem with $\mathcal{S}_{\max}$'s answer for ANDDUPLICATE is that it inherits the same overestimated synergy from AND (discussed in Section 4.3).

5 Synergistic mutual information

We are all familiar with the English expression describing synergy as the whole being greater than the "sum of its parts". Although this informal adage, formalized by WMS synergy, captures the intuition behind synergy, we saw that WholeMinusSum "double-counts" whenever there is duplication (redundancy) among the parts. A mathematically correct adage should change "sum" to "union"—meaning synergy occurs when the whole is greater than the *union* of its parts. Summing adds duplicate information multiple times, whereas union adds duplicate information only once. The union of the parts never exceeds the sum.

This guiding intuition of "whole minus union" leads us to a novel measure entitled *synergistic mutual information*, denoted $\mathcal{S}\left(\{X_1, \ldots, X_n\} : Y\right)$, or $\mathcal{S}\left(\mathbf{X} : Y\right)$, as the mutual information in the whole that is not in the union of its parts.

Unfortunately, there's no measure of 'union-information" in contemporary information theory. We introduce a novel technique, derived from [26, 27], for defining the union information among n predictors. We numerically compute the union information by passing Y through a channel (particularly, a square transition matrix) that preserves only the bits that are specified by singleton predictors. This is achieved like so,

$$\mathrm{I}_{\cup}\left(\{X_1, \ldots, X_n\} : Y\right) \equiv \min_{\Pr(Y'|Y)} \mathrm{I}\left(X_{1 \ldots n} : Y'\right) \tag{14}$$
$$\text{subject to} \quad X_{1 \ldots n} \to Y \to Y'$$
$$\mathrm{I}\left(X_i : Y'\right) = \mathrm{I}(X_i : Y) \ \forall i \,.$$

The constraint $X_{1 \ldots n} \to Y \to Y'$ is a Markov chain placing Y between $X_{1 \ldots n}$ and Y'. This Markov chain ensures that *all information* between $X_{1 \ldots n}$ and Y' is also between $X_{1 \ldots n}$ and Y—thus $\mathrm{I}\left(X_{1 \ldots n} : Y'\right) \leq \mathrm{I}(X_{1 \ldots n} : Y)$. An equivalent way of conceptualizing this Markov chain is that it forces the joint distribution $\Pr\left(x_{1 \ldots n}, y, y'\right) = \Pr(x_{1 \ldots n}, y) \Pr\left(y'|y\right)$. Without

X_1 X_2 X_3	Y	
0 0 0	0	$1/4$
0 1 0	0	$1/4$
1 0 1	0	$1/4$
1 1 1	1	$1/4$

(a) $\Pr(x_1, x_2, x_3, y)$

(b) circuit diagram

(c) PI-diagram
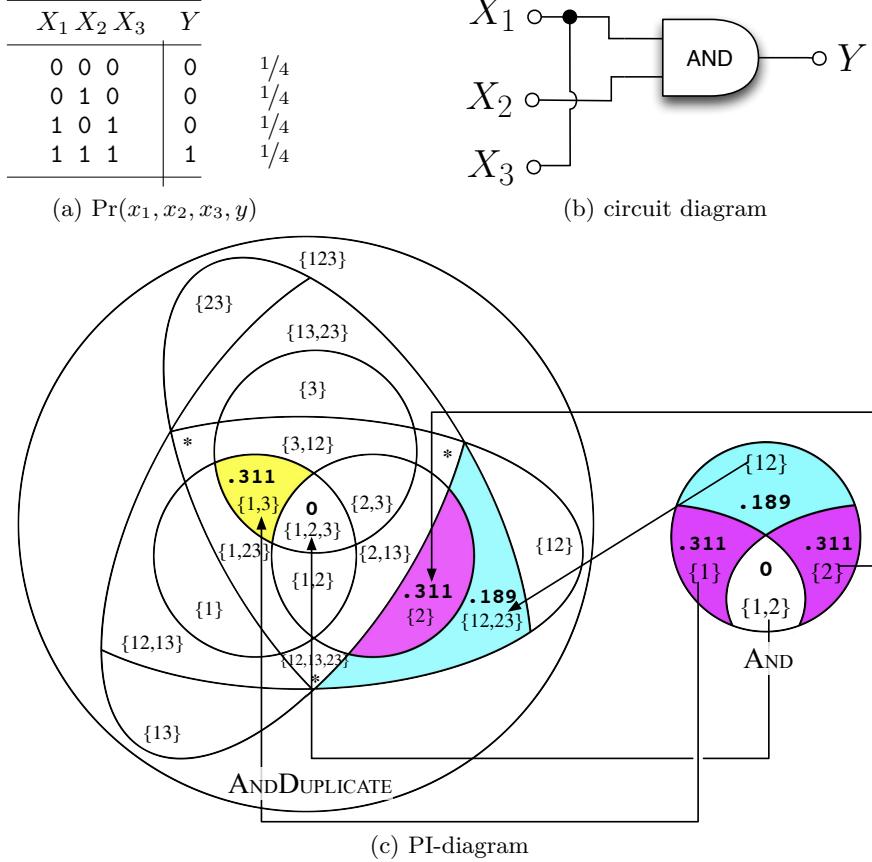
Figure 10: Example ANDDUPLICATE. The total mutual information is the same as in AND, $I(X_1 X_2 : Y) = I(X_1 X_2 X_3 : Y) = 0.811$ bits. Every PI-region in example AND (Figure 9c) maps to a PI-region in ANDDUPLICATE. At 0.189 bits, the intuitive synergistic information is unchanged from AND.

another constraint, $\min_{X_{1 \ldots n} \to Y \to Y'} I(X_{1 \ldots n} : Y')$ is trivially found to be zero bits. This is because simply setting Y' to a constant would have $H(Y') = 0$ bits, thus necessitating $I(X_{1 \ldots n} : Y') = 0$ bits. Therefore we must also explicitly state which bits of Y we wish to retain—those bits originating from the singletons $\{X_1, \ldots, X_n\}$.

Taken together these constraints ensure that: (1) All information between $X_{1 \ldots n}$ and Y' is also between $X_{1 \ldots n}$ and Y; (2) $I(X_{1 \ldots n} : Y')$ only contains bits that the constraints explicitly preserve—those between $I(X_i : Y)$ $\forall i$. Finally, we prove that a minimum of eq. (14) always exists because setting $\Pr(Y'|Y)$ to the identity matrix satisfies all constraints.[8]

Unfortunately we currently have no analytic way to calculate $I_\cup$ (eq. (14)). In practice we use MATLAB to perform gradient descent optimization using the function `fmincon`. We have explored some of the properties of the minimization in eq. (14). First, because of the constraint $I(X_i : Y') = I(X_i : Y)$ $\forall i$, the minimization is unfortunately not convex—therefore there's no straightforward way to verify we've found the minimum. However, for $n = 2$ the minimization procedure can be rewritten as a convex minimization [25, 27]. Second, the minimizing matrix $\Pr(Y'|Y)$ is not unique. We are exploring analytic solutions for $I_\cup$. Our union-information measure satisfies all desired properties for a union-measure from [24],

---

[8]Because of the specific nature of the minimization removing the Markov constraint $X_{1 \ldots n} \to Y \to Y'$ doesn't alter the res ult of eq. (14), but the resulting equation becomes less intuitive and harder to numerically compute.

non-negativity, idempotency, symmetry among X_i's, nondecreasing with additional X_i's, and equal to the entropy $\mathrm{H}(X_{1\dots n})$ when $Y = X_{1\dots n}$.

Once the union information is computed, we define the synergistic mutual information among the n predictors as,

$$\mathcal{S}\left(\{X_1, \dots, X_n\} : Y\right) \equiv \mathrm{I}(X_{1\dots n} : Y) - \mathrm{I}_\cup\left(\{X_1, \dots, X_n\} : Y\right) \ . \tag{15}$$

Synergistic mutual information quantifies the total "informational work" *only coalitions* perform in reducing the uncertainty of Y. Pleasingly, synergistic mutual information measure is bounded[9] by the WholeMinusSum synergy (which underestimates the intuitive synergy) and $\mathcal{S}_{\max}$ (which overestimates intuitive synergy),

$$\max\left[0, \mathrm{WMS}\left(\mathbf{X} : Y\right)\right] \leq \mathcal{S}\left(\mathbf{X} : Y\right) \leq \mathcal{S}_{\max}\left(\mathbf{X} : Y\right) \leq \mathrm{I}(X_{1\dots n} : Y) \ . \tag{16}$$

Synergistic mutual information can be made state-dependent for a particular state $y \in Y$, i.e. $\mathcal{S}(\mathbf{X} : Y = y)$, Appendix C has the details.

If there is no redundant information among the parts, which is guaranteed by the simple condition $\sum_{i=1}^n \mathrm{I}\left(X_i : X_{1\dots n \setminus i}\right) = 0$,[10] the synergistic mutual information is equal to WholeMinusSum, $\mathcal{S}\left(\mathbf{X} : Y\right) = \mathrm{WMS}\left(\mathbf{X} : Y\right)$.

Conditional dependence among predictors $\mathbf{X}$, $\Pr\left(X_{1\dots n}|Y\right) \neq \prod_{i=1}^n \Pr\left(X_i|Y\right)$ is necessary but not sufficient for set $\mathbf{X}$ to have synergistic information about Y.

6   Applying the measures to our examples

Table 1 summarizes the results of all four measures applied to our examples.

RDN (Figure 2). There is exactly one bit of redundant information and all measures reach their intended answer.

UNQ (Figure 3). $\mathcal{S}_{\max}$'s miscategorization of unique information as synergistic information reveals itself. Intuitively, there are two bits of unique information and no synergy. However, $\mathcal{S}_{\max}$ reports one bit of synergistic information.

XOR (Figure 4). There is one bit of synergistic information and nothing more. All measures reach the expected answer of 1 bit.

XORDUPLICATE (Figure 5). Target Y is specified by the coalition $X_1 X_2$ as well as by the coalition $X_3 X_2$, thus $\mathrm{I}(X_1 X_2 : Y) = \mathrm{I}(X_3 X_2 : Y) = \mathrm{H}(Y) = 1$ bit. All measures reach the expected answer of 1 bit.

XORLOSES (Figure 6). Target Y is fully specified by the coalition $X_1 X_2$ as well as by the singleton X_3, thus $\mathrm{I}(X_1 X_2 : Y) = \mathrm{I}(X_3 : Y) = \mathrm{H}(Y) = 1$ bit. Together this means there is one bit of redundancy between the coalition $X_1 X_2$ and the singleton X_3 as denoted by the $+1$ in PI-region $\{3, 12\}$. All measures account for this redundancy and reach the expected answer of 0 bits.

RDNXOR (Figure 8). This example has one bit of synergy as well as one bit of redundancy. In accordance with Figure 7a, WholeMinusSum measures *synergy minus redundancy* to calculate $1 - 1 = 0$ bits. On the other hand, $\mathcal{S}_{\max}$, $\Delta\mathrm{I}$, and $\mathcal{S}$ are not mislead by the co-existance of synergy and redundancy and correctly report 1 bit of synergistic information.

AND (Figure 9). This example is a simple case where correlational importance, $\Delta\mathrm{I}(\mathbf{X}; Y)$, disagrees with the intuitive value for synergy. The WholeMinusSum synergy—the *lowerbound* on the intuitive synergy—is 0.189 bits, yet $\Delta\mathrm{I}(\mathbf{X}; Y) = 0.104$ bits. Furthermore, just as in example UNQ, $\mathcal{S}_{\max}$ again miscategorizes the second unique information as synergistic to overestimate the synergy arriving at $0.189 + 0.311 = 0.5$ bits.

---

[9]Proven in Appendix D.2.

[10]This condition is much looser than full mutual independence among predictors.

Example	$\mathcal{S}_{\max}$	WMS	Δ I	$\mathcal{S}$
RDN	0	−1	0	0
UNQ	1	0	0	0
XOR	1	1	1	1
XORDUPLICATE	1	1	1	1
XORLOSES	0	0	0	0
RDNXOR	1	0	1	1
AND	$1/2$	0.189	0.104	0.189
ANDDUPLICATE	$1/2$	−0.123	0.038	0.189
XORMULTICOAL	1	1	1	1
RDNUNQXOR	2	0	1	1
PARITYRDNRDN	1	−3	1	1

Table 1: Synergy measures for our examples. Answers conflicting with the intuitive synergistic information are in red. All examples exploit special cases to analytically compute $\mathcal{S}$. Both analytically and numerically our measure $\mathcal{S}$ reaches the intuitive answer for every example.

ANDDUPLICATE (Figure 10). This example shows how the measures respond to duplicating a predictor for example AND. As first demonstrated in example XORDUPLICATE, intuitive synergistic information is unchanged when duplicating a predictor. However, both WholeMinusSum and Δ I conflict with this intuition to *decrease* from AND to ANDDUPLICATE. In contrast, measures $\mathcal{S}_{\max}$ and $\mathcal{S}$ are invariant when duplicating predictors.

The three final examples XORMULTICOAL, RDNUNQXOR, and PARITYRDNRDN aren't essential for understanding this paper and are discussed in Appendix A.

7 Discussion

Fundamentally, we assert that synergy quantifies how much a whole is exceeds the *union* of its parts. Considering synergy as the whole minus the *sum* of its parts inadvertently "double-subtracts" redundancies, thus *underestimating* synergy. Within information theory, PI-diagrams, a generalization of Venn diagrams introduced in [6], are immensely helpful in improving one's intuition for synergy.

Table 1 shows that no prior measure quantifies the intuitive notion of synergistic information in all cases. In fact, no prior measure consistently matches intuition even for $n = 2$. To summarize,

1. $I_{\max}$ synergy, $\mathcal{S}_{\max}$, overestimates the intutive synergy when two or more predictors have unique information about the target (e.g. UNQ).

2. WholeMinusSum synergy, WMS, inadvertently double-subtracts redundancies and thus underestimates the intuitive synergy (e.g. RDNXOR). Duplicating predictors turns unique information into redundant information thereby decreasing WholeMinusSum synergy.

3. Correlational importance, Δ I, is not bounded by the Shannon mutual information. Duplicating predictors often decreases correlational importance (e.g. ANDDUPLICATE). Altogether, Δ I does not quantify the intuitive synergistic information (nor was it intended to).

We demonstrate by examples (e.g. RDNXOR and RDNUNQXOR in Appendix A) that a single state can simutaneously carry redundant, unique, and synergistic information. This fact is underappreciated in the current literature. Prior work often implicitly assumed that these three types of information cannot coexist in a single state.

We introduce an implicit analytical expression for synergistic mutual information (eq. (15)). Unfortunately our expression is not easily computable, and until we have an explicit analytic

derivation of the union information the best one can do is compute synergistic mutual information via numerical optimization techniques. Along with our examples, we consider our introduction of a necessary and sufficient criteria for the union information (eq. (14)) our primary contribution to the literature.

We believe that our measure of synergy, *synergistic mutual information*, will be important in untangling informational relationships among the heavily interconnected molecular, genomic and neuronal networks found in evolved biological systems characterized by a high degree of robustness and redundancy.

Acknowledgments

References

[1] Narayanan NS, Kimchi EY, Laubach M (2005) Redundancy and synergy of neuronal ensembles in motor cortex. The Journal of Neuroscience 25: 4207-4216.

[2] Balduzzi D, Tononi G (2008) Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS Computational Biology 4: e1000091.

[3] Schneidman E, Bialek W, II MB (2003) Synergy, redundancy, and independence in population codes. Journal of Neuroscience 23: 11539–53.

[4] Bell AJ (2003) The co-information lattice. In: Amari S, Cichocki A, Makino S, Murata N, editors, Fifth International Workshop on Independent Component Analysis and Blind Signal Separation. Springer.

[5] Nirenberg S, Carcieri SM, Jacobs AL, Latham PE (2001) Retinal ganglion cells act largely as independent encoders. Nature 411: 698–701.

[6] Williams PL, Beer RD (2010) Nonnegative decomposition of multivariate information. CoRR abs/1004.2515.

[7] White D, Rabago-Smith M (2011) Genotype-phenotype associations and human eye color. Journal of Human Genetics 56: 5–7.

[8] Schneidman E, Still S, Berry MJ, Bialek W (2003) Network information and connected correlations. Phys Rev Lett 91: 238701-238705.

[9] Anastassiou D (2007) Computational analysis of the synergy among multiple interacting genes. Molecular Systems Biology 3: 83.

[10] Cover TM, Thomas JA (1991) Elements of Information Theory. New York, NY: John Wiley.

[11] Weisstein EW (2011). Antichain. `http://mathworld.wolfram.com/Antichain.html`.

[12] Comtet L (1998) Advanced Combinatorics: The Art of Finite and Infinite Expansions. Dordrecht, Netherlands: Reidel, 271–273 pp.

[13] DeWeese MR, Meister M (1999) How to measure the information gained from one symbol. Network 10: 325-340.

[14] Gawne TJ, Richmond BJ (1993) How independent are the messages carried by adjacent inferior temporal cortical neurons? Journal of Neuroscience 13: 2758-71.

[15] Gat I, Tishby N (1999) Synergy and redundancy among brain cells of behaving monkeys. In: Advances in Neural Information Proceedings systems. MIT Press, pp. 465–471.

[16] Chechik G, Globerson A, Anderson MJ, Young ED, Nelken I, et al. (2002) Group redundancy measures reveal redundancy reduction in the auditory pathway. In: Dietterich TG, Becker S, Ghahramani Z, editors, NIPS 2002. Cambridge, MA: MIT Press, pp. 173–180.

[17] Han TS (1978) Nonnegative entropy measures of multivariate symmetric correlations. Information and Control 36: 133–156.

[18] Panzeri S, Treves A, Schultz S, Rolls ET (1999) On decoding the responses of a population of neurons from short time windows. Neural Comput 11: 1553–1577.

[19] Nirenberg S, Latham PE (2003) Decoding neuronal spike trains: How important are correlations? Proceedings of the National Academy of Sciences 100: 7348–7353.

[20] Pola G, Thiele A, Hoffmann KP, Panzeri S (2003) An exact method to quantify the information transmitted by different mechanisms of correlational coding. Network 14: 35–60.

[21] Latham PE, Nirenberg S (2005) Synergy, redundancy, and independence in population codes, revisited. Journal of Neuroscience 25: 5195-5206.

[22] Lei W, Xu G, Chen B (2010) The common information of n dependent random variables. Forty-Eighth Annual Allerton Conference on Communication, Control, and Computing abs/1010.3613: 836–843.

[23] Kamath S, Anantharam V (2010) A new dual to the gács-körner common information defined via the gray-wyner system. Forty-Eighth Annual Allerton Conference on Communication, Control, and Computing : 1340-46.

[24] Harder M, Salge C, Polani D (2012) A bivariate measure of redundant information. CoRR abs/1207.2080.

[25] Griffith V (2012) Bivariate redundancy and synergy, or: understanding conditional mutual information. in press .

[26] Maurer UM, Wolf S (1999) Unconditionally secure key agreement and the intrinsic conditional information. IEEE Transactions on Information Theory 45: 499-514.

[27] Christandl M, Renner R, Wolf S (2003) A property of the intrinsic mutual information. In: Proceedings of the IEEE International Symposium on Information Theory. p. 258. doi:10.1109/ISIT.2003.1228272.

Appendix

A Three extra examples

For the reader's intellectual pleasure, we include three more sophisticated examples: XOR-MULTICOAL, RDNUNQXOR, and PARITYRDNRDN. Example RDNUNQXOR extends example RDNXOR to demonstrate redundant, unique, and synergistic information for every state $y \in Y$. Example PARITYRDNRDN illustrates how for $n > 2$, WholeMinusSum synergy subtracts redundancies multiple times.

X_1	X_2	X_3	Y	
ab	ac	bc	0	$1/8$
AB	Ac	Bc	0	$1/8$
Ab	AC	bC	0	$1/8$
aB	aC	BC	0	$1/8$
Ab	Ac	bc	1	$1/8$
aB	ac	Bc	1	$1/8$
ab	aC	bC	1	$1/8$
AB	AC	BC	1	$1/8$

(a) $\Pr(x_1, x_2, x_3, y)$
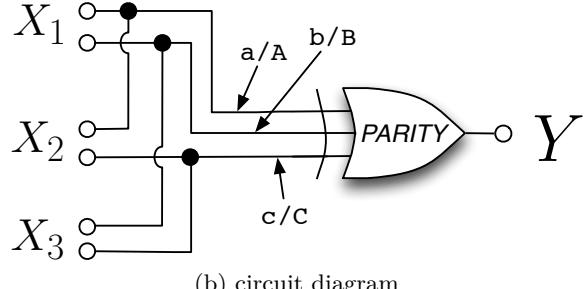
(b) circuit diagram
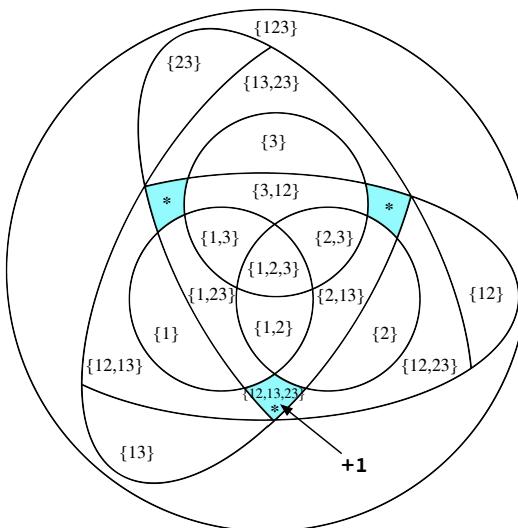
(c) PI-diagram

Figure 11: Example XORMULTICOAL demonstrates how the same information can be specified by multiple coalitions. In XORMULTICOAL the target Y has one bit of uncertainty, $\mathrm{H}(Y) = 1$ bit, and Y is the *parity* of three incoming wires. Just as the output of XOR is specified only after knowing the state of both inputs, the output of XORMULTICOAL is specified only after knowing the state of all three wires. Each predictor is distinct and has access to two of the three incoming wires. For example, predictor X_1 has access to the a/A and b/B wires, X_2 has access to the a/A and c/C wires, and X_3 has access to the b/B and c/C wires. Although no single predictor specifies Y, any coalition of two predictors has access to all three wires and fully specifies Y, $\mathrm{I}(X_1 X_2{:}Y) = \mathrm{I}(X_1 X_3{:}Y) = \mathrm{I}(X_2 X_3{:}Y) = \mathrm{H}(Y) = 1$ bit. In the PI-diagram this puts one bit in PI-region $\{12, 13, 23\}$ and zero everywhere else. The amount of synergistic information is the same as XOR, and all measures reach the expected answer of 1 bit.

X_1	X_2	Y	
ra0	rb0	rab0	$^1/_{32}$
ra0	rb1	rab1	$^1/_{32}$
ra1	rb0	rab1	$^1/_{32}$
ra1	rb1	rab0	$^1/_{32}$
ra0	rB0	raB0	$^1/_{32}$
ra0	rB1	raB1	$^1/_{32}$
ra1	rB0	raB1	$^1/_{32}$
ra1	rB1	raB0	$^1/_{32}$
rA0	rb0	rAb0	$^1/_{32}$
rA0	rb1	rAb1	$^1/_{32}$
rA1	rb0	rAb1	$^1/_{32}$
rA1	rb1	rAb0	$^1/_{32}$
rA0	rB0	rAB0	$^1/_{32}$
rA0	rB1	rAB1	$^1/_{32}$
rA1	rB0	rAB1	$^1/_{32}$
rA1	rB1	rAB0	$^1/_{32}$

X_1	X_2	Y	
Ra0	Rb0	Rab0	$^1/_{32}$
Ra0	Rb1	Rab1	$^1/_{32}$
Ra1	Rb0	Rab1	$^1/_{32}$
Ra1	Rb1	Rab0	$^1/_{32}$
Ra0	RB0	RaB0	$^1/_{32}$
Ra0	RB1	RaB1	$^1/_{32}$
Ra1	RB0	RaB1	$^1/_{32}$
Ra1	RB1	RaB0	$^1/_{32}$
RA0	Rb0	RAb0	$^1/_{32}$
RA0	Rb1	RAb1	$^1/_{32}$
RA1	Rb0	RAb1	$^1/_{32}$
RA1	Rb1	RAb0	$^1/_{32}$
RA0	RB0	RAB0	$^1/_{32}$
RA0	RB1	RAB1	$^1/_{32}$
RA1	RB0	RAB1	$^1/_{32}$
RA1	RB1	RAB0	$^1/_{32}$

(a) $\Pr(x_1, x_2, y)$

(b) circuit diagram

(c) PI-diagram
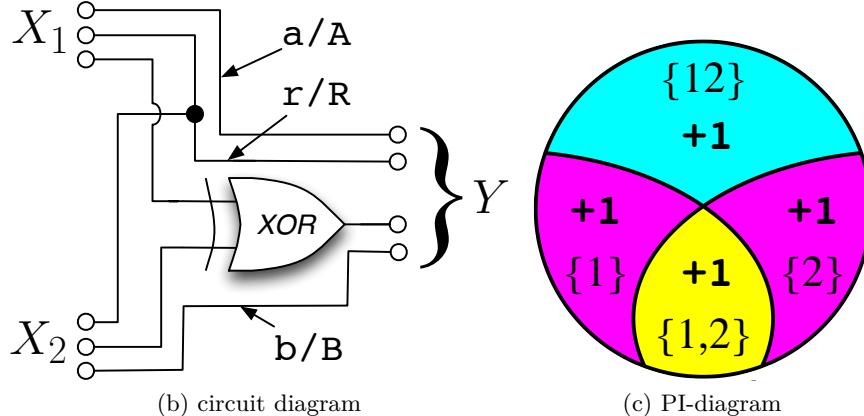
Figure 12: Example RDNUNQXOR weaves examples RDN, UNQ, and XOR into one. $I(X_1 X_2 : Y) = H(Y) = 4$ bits. This example is pleasing because it puts exactly one bit in every PI-region.

X_1	X_2	X_3	Y	
ab0	ab0	ab0	ab0	$1/32$
ab0	ab0	ab1	ab1	$1/32$
ab0	ab1	ab0	ab1	$1/32$
ab0	ab1	ab1	ab0	$1/32$
ab1	ab0	ab0	ab1	$1/32$
ab1	ab0	ab1	ab0	$1/32$
ab1	ab1	ab0	ab0	$1/32$
ab1	ab1	ab1	ab1	$1/32$
aB0	aB0	aB0	aB0	$1/32$
aB0	aB0	aB1	aB1	$1/32$
aB0	aB1	aB0	aB1	$1/32$
aB0	aB1	aB1	aB0	$1/32$
aB1	aB0	aB0	aB1	$1/32$
aB1	aB0	aB1	aB0	$1/32$
aB1	aB1	aB0	aB0	$1/32$
aB1	aB1	aB1	aB1	$1/32$

X_1	X_2	X_3	Y	
Ab0	Ab0	Ab0	Ab0	$1/32$
Ab0	Ab0	Ab1	Ab1	$1/32$
Ab0	Ab1	Ab0	Ab1	$1/32$
Ab0	Ab1	Ab1	Ab0	$1/32$
Ab1	Ab0	Ab0	Ab1	$1/32$
Ab1	Ab0	Ab1	Ab0	$1/32$
Ab1	Ab1	Ab0	Ab0	$1/32$
Ab1	Ab1	Ab1	Ab1	$1/32$
AB0	AB0	AB0	AB0	$1/32$
AB0	AB0	AB1	AB1	$1/32$
AB0	AB1	AB0	AB1	$1/32$
AB0	AB1	AB1	AB0	$1/32$
AB1	AB0	AB0	AB1	$1/32$
AB1	AB0	AB1	AB0	$1/32$
AB1	AB1	AB0	AB0	$1/32$
AB1	AB1	AB1	AB1	$1/32$

(a) $\Pr(x_1, x_2, x_3, y)$

(b) circuit diagram

(c) PI-diagram

Figure 13: Example PARITYRDNRDN has three predictors. The target Y has three bits of uncertainty—H$(Y) = 3$. Examining any singleton predictor specifies the letters in Y (ab/aB/Ba/AB), I$(X_i \!:\! Y) = 2\ \forall i$, making two bits of redundant information. Y's third and final bit (digit 0/1) is the parity of the digits of the three predictors and accordingly is specified only by the triplet coalition $X_1 X_2 X_3$, making one bit of synergy. This example has two bits of maximum redundancy and one bit of synergy. I$(X_1 X_2 X_3 \!:\! Y) = $ H$(Y) = 3$ bits.

B   Algebraic simplification of ΔI

Prior literature [5, 19–21] defines $\Delta\,\mathrm{I}\,(\mathbf{X};Y)$ as,

$$\Delta\,\mathrm{I}\,(\mathbf{X};Y) \quad\equiv\quad \mathrm{D}_{\mathrm{KL}}\Big[\mathrm{Pr}\big(Y|X_{1\ldots n}\big)\Big\|\mathrm{Pr}_{\mathrm{ind}}\,(Y|\mathbf{X})\Big] \tag{17}$$

$$=\quad \mathbb{E}_{\mathbf{X}}\,\mathrm{D}_{\mathrm{KL}}\Big[\mathrm{Pr}\big(Y|\mathbf{x}\big)\Big\|\mathrm{Pr}_{\mathrm{ind}}(Y|\mathbf{x})\Big] \tag{18}$$

$$=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(y|\mathbf{x}\big)}{\mathrm{Pr}_{\mathrm{ind}}(y|\mathbf{x})}\;. \tag{19}$$

Where,

$$\mathrm{Pr}_{\mathrm{ind}}(Y=y|\mathbf{X}=\mathbf{x}) \quad\equiv\quad \frac{\mathrm{Pr}(y)\,\mathrm{Pr}_{\mathrm{ind}}(\mathbf{X}=\mathbf{x}|Y=y)}{\mathrm{Pr}_{\mathrm{ind}}(\mathbf{X}=\mathbf{x})} \tag{20}$$

$$=\quad \frac{\mathrm{Pr}(y)\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)}{\mathrm{Pr}_{\mathrm{ind}}(\mathbf{x})} \tag{21}$$

$$\mathrm{Pr}_{\mathrm{ind}}(\mathbf{X}=\mathbf{x}) \quad\equiv\quad \mathbb{E}_{Y}\left[\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)\right] \tag{22}$$

$$=\quad \sum_{y\in Y}\mathrm{Pr}(Y=y)\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big) \tag{23}$$

The definition of $\Delta\,\mathrm{I}$ (eq. (17)) reduces to,

$$\Delta\,\mathrm{I}\,(\mathbf{X};Y) \quad=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(y|\mathbf{x}\big)}{\mathrm{Pr}_{\mathrm{ind}}(y|\mathbf{x})} \tag{24}$$

$$=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(y|\mathbf{x}\big)\,\mathrm{Pr}_{\mathrm{ind}}(\mathbf{x})}{\mathrm{Pr}(y)\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)} \tag{25}$$

$$=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)}\frac{\mathrm{Pr}_{\mathrm{ind}}(\mathbf{x})}{\mathrm{Pr}(\mathbf{x})} \tag{26}$$

$$=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)}+\sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}_{\mathrm{ind}}(\mathbf{x})}{\mathrm{Pr}(\mathbf{x})} \tag{27}$$

$$=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)}+\sum_{\mathbf{x}\in\mathbf{X}}\mathrm{Pr}(\mathbf{x})\log\frac{\mathrm{Pr}_{\mathrm{ind}}(\mathbf{x})}{\mathrm{Pr}(\mathbf{x})} \tag{28}$$

$$=\quad \sum_{\mathbf{x},y\in\mathbf{X},Y}\mathrm{Pr}(\mathbf{x},y)\log\frac{\mathrm{Pr}\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\mathrm{Pr}\big(x_i|y\big)}-\sum_{\mathbf{x}\in\mathbf{X}}\mathrm{Pr}(\mathbf{x})\log\frac{\mathrm{Pr}(\mathbf{x})}{\mathrm{Pr}_{\mathrm{ind}}(\mathbf{x})} \tag{29}$$

$$=\quad \mathrm{D}_{\mathrm{KL}}\left[\mathrm{Pr}\big(X_{1\ldots n}|Y\big)\Big\|\prod_{i=1}^{n}\mathrm{Pr}\big(X_i|Y\big)\right]-\mathrm{D}_{\mathrm{KL}}\big[\mathrm{Pr}(X_{1\ldots n})\big\|\mathrm{Pr}_{\mathrm{ind}}(\mathbf{X})\big] \tag{30}$$

$$=\quad \mathrm{TC}\,(X_1;\cdots;X_n|Y)-\mathrm{D}_{\mathrm{KL}}\big[\mathrm{Pr}(X_{1\ldots n})\big\|\mathrm{Pr}_{\mathrm{ind}}(\mathbf{X})\big] \tag{31}$$

$$=\quad \mathrm{TC}\,(X_1;\cdots;X_n|Y)-\mathrm{D}_{\mathrm{KL}}\left[\mathrm{Pr}(X_{1\ldots n})\Big\|\sum_{y\in Y}\mathrm{Pr}(y)\prod_{i=1}^{n}\mathrm{Pr}\big(X_i|y\big)\right]\,. \tag{32}$$

where $\mathrm{TC}\,(X_1;\cdots;X_n|Y)$ is the conditional total correlation among the predictors given Y.

C State-dependent synergistic mutual information

We define synergistic mutual information, $\mathcal{S}$, for a particular state of the target $y \in Y$ as,

$$\mathcal{S}\left(\{X_1,\ldots,X_n\} : Y = y\right) \equiv \mathrm{I}(X_{1\ldots n} : Y = y) - \mathrm{I}_\cup\left(\{X_1,\ldots,X_n\} : Y = y\right) , \qquad (33)$$

where the state-dependent union-information ($\mathrm{I}_\cup$) is,

$$\mathrm{I}_\cup\left(\{X_1,\ldots,X_n\} : Y = y\right) \equiv \min_{\Pr(y'|Y)} \mathrm{I}\left(X_{1\ldots n} : Y' = y'\right) \qquad (34)$$

$$\text{subject to} \quad X_{1\ldots n} \to Y \to y'$$

$$\mathrm{I}\left(X_i : Y' = y'\right) = \mathrm{I}(X_i : Y = y) \ \forall i .$$

The state-dependent mutual information, e.g. $\mathrm{I}(X_i : Y = y)$, is [13]'s "specific-surprise",

$$\mathrm{I}(X_i : Y = y) \quad \equiv \quad \mathrm{D_{KL}}\Big[\Pr\big(X_i|y\big) \,\big\|\, \Pr(X_i)\Big] \qquad (35)$$

$$= \quad \sum_{x_i \in X_i} \Pr\big(x_i|y\big) \log \frac{\Pr(x_i, y)}{\Pr(x_i)\Pr(y)} . \qquad (36)$$

D Essential proofs

These proofs underpin our essential claims about our introduced measure, synergistic mutual information.

D.1 Proof of equivalence to the Maurer method for $n = 2$

First, an initial proof that,

$$\mathrm{I}(X_1 : X_2|Y) = \mathrm{I}(X_1 X_2 : Y) + \mathrm{I}(X_1 : X_2) - \mathrm{I}(X_1 : Y) - \mathrm{I}(X_2 : Y) .$$

Proof.

$$
\begin{aligned}
\mathrm{I}(X_1 X_2 : Y) - \mathrm{I}(X_1 : Y) - \mathrm{I}(X_2 : Y) \quad &= \quad \mathrm{H}(X_1 X_2) - \mathrm{H}\big(X_1 X_2|Y\big) - \mathrm{H}(X_1) + \mathrm{H}\big(X_1|Y\big) - \mathrm{H}(X_2) + \mathrm{H}\big(X_2|Y\big) \\
&= \quad \mathrm{I}\big(X_1 : X_2|Y\big) + \mathrm{H}(X_1 X_2) - \mathrm{H}(X_1) - \mathrm{H}(X_2) \qquad (37) \\
&= \quad \mathrm{I}\big(X_1 : X_2|Y\big) - \mathrm{I}(X_1 : X_2) \qquad (38) \\
\mathrm{I}\big(X_1 : X_2|Y\big) \quad &= \quad \mathrm{I}(X_1 X_2 : Y) + \mathrm{I}(X_1 : X_2) - \mathrm{I}(X_1 : Y) - \mathrm{I}(X_2 : Y) . \qquad (39)
\end{aligned}
$$

$\square$

Now we prove that for $n = 2$ the Maurer-method for computing synergy is equivalent to our method for computing synergy. We show that,

$$
\begin{aligned}
\mathrm{I}\big(X_1 : X_2|Y\big) - \min_{X_1 X_2 \to Y \to Y'} \mathrm{I}\big(X_1 : X_2|Y'\big) \quad &= \quad \mathcal{S}\left(\{X_1, X_2\} : Y\right) \qquad (40) \\
&= \quad \mathrm{I}(X_1 X_2 : Y) - \min_{\substack{X_1 X_2 \to Y \to Y' \\ \mathrm{I}(X_1:Y')=\mathrm{I}(X_1:Y) \\ \mathrm{I}(X_2:Y')=\mathrm{I}(X_2:Y)}} \mathrm{I}\big(X_1 X_2 : Y'\big) .
\end{aligned}
$$

Proof.

$$\mathrm{I}\big(X_1\!:\!X_2|Y\big) - \min_{X_1 X_2 \to Y \to Y'} \underbrace{\mathrm{I}\big(X_1\!:\!X_2\big|Y'\big)}_{\text{expand per eq. (39)}} \tag{41}$$

$$= \mathrm{I}\big(X_1\!:\!X_2|Y\big) - \min_{X_1 X_2 \to Y \to Y'} \Big[\mathrm{I}\big(X_1 X_2\!:\!Y'\big) + \mathrm{I}\big(X_1\!:\!X_2\big) - \mathrm{I}\big(X_1\!:\!Y'\big) - \mathrm{I}\big(X_2\!:\!Y'\big)\Big] \tag{42}$$

$$= \underbrace{\mathrm{I}\big(X_1\!:\!X_2|Y\big)}_{\text{expand per eq. (39)}} - \mathrm{I}\big(X_1\!:\!X_2\big) - \min_{X_1 X_2 \to Y \to Y'} \Big[\mathrm{I}\big(X_1 X_2\!:\!Y'\big) - \mathrm{I}\big(X_1\!:\!Y'\big) - \mathrm{I}\big(X_2\!:\!Y'\big)\Big] \tag{43}$$

$$= \mathrm{I}(X_1 X_2\!:\!Y) - \mathrm{I}(X_1\!:\!Y) - \mathrm{I}(X_2\!:\!Y) - \min_{X_1 X_2 \to Y \to Y'} \left[\underbrace{\mathrm{I}\big(X_1 X_2\!:\!Y'\big) - \mathrm{I}\big(X_1\!:\!Y'\big) - \mathrm{I}\big(X_2\!:\!Y'\big)}_{\text{decompose into PI-regions}}\right].$$

We now decompose $\mathrm{I}\big(X_1 X_2\!:\!Y'\big) - \mathrm{I}\big(X_1\!:\!Y'\big) - \mathrm{I}\big(X_2\!:\!Y'\big)$ into PI-regions.

- $\mathrm{I}\big(X_1 X_2\!:\!Y'\big)$ is composed of PI-regions: {12}, {1}, {2}, and {1,2}.
- $\mathrm{I}\big(X_1\!:\!Y'\big)$ is composed of PI-regions {1} and {1,2}.
- $\mathrm{I}\big(X_2\!:\!Y'\big)$ is composed of PI-regions {2} and {1,2}.

Thus the difference $\mathrm{I}\big(X_1 X_2\!:\!Y'\big) - \mathrm{I}\big(X_1\!:\!Y'\big) - \mathrm{I}\big(X_2\!:\!Y'\big)$ is PI-regions $\{12\} - \{1,2\}$.

$$= \mathrm{I}(X_1 X_2\!:\!Y) - \mathrm{I}(X_1\!:\!Y) - \mathrm{I}(X_2\!:\!Y) - \min_{X_1 X_2 \to Y \to Y'} \left[\underbrace{\mathrm{I}\big(X_1 X_2\!:\!Y'\big) - \mathrm{I}\big(X_1\!:\!Y'\big) - \mathrm{I}\big(X_2\!:\!Y'\big)}_{\text{PI-regions: } \{12\} - \{1,2\}}\right]. \tag{44}$$

As the minimum of eq. (44) is the synergy (PI-region {12}) minus the redundancy (PI-region {1,2}), we can add any constraints we wish to the minimization $\min_{X_1 X_2 \to Y \to Y'}$ that do not increase the synergy or decrease the redundancy. We choose to add the constraints $\mathrm{I}\big(X_1\!:\!Y'\big) = \mathrm{I}(X_1\!:\!Y)$ and $\mathrm{I}\big(X_2\!:\!Y'\big) = \mathrm{I}(X_2\!:\!Y)$. This gives us,

$$\mathrm{I}(X_1 X_2\!:\!Y) - \mathrm{I}(X_1\!:\!Y) - \mathrm{I}(X_2\!:\!Y) - \min_{\substack{X_1 X_2 \to Y \to Y' \\ I(X_1:Y')=I(X_1:Y) \\ I(X_2:Y')=I(X_2:Y)}} \left[\mathrm{I}\big(X_1 X_2\!:\!Y'\big) - \underbrace{\mathrm{I}\big(X_1\!:\!Y'\big)}_{=I(X_1:Y)} - \underbrace{\mathrm{I}\big(X_2\!:\!Y'\big)}_{=I(X_2:Y)}\right]$$

$$= \mathrm{I}(X_1 X_2\!:\!Y) - \mathrm{I}(X_1\!:\!Y) - \mathrm{I}(X_2\!:\!Y) + \mathrm{I}(X_1\!:\!Y) + \mathrm{I}(X_2\!:\!Y) \min_{\substack{X_1 X_2 \to Y \to Y' \\ I(X_1:Y')=I(X_1:Y) \\ I(X_2:Y')=I(X_2:Y)}} \mathrm{I}\big(X_1 X_2\!:\!Y'\big) \tag{45}$$

$$= \mathrm{I}(X_1 X_2\!:\!Y) - \min_{\substack{X_1 X_2 \to Y \to Y' \\ I(X_1:Y')=I(X_1:Y) \\ I(X_2:Y')=I(X_2:Y)}} \mathrm{I}\big(X_1 X_2\!:\!Y'\big) \tag{46}$$

$$= \mathcal{S}\big(\{X_1, X_2\} : Y\big). \tag{47}$$

And the proof of eq. (40) is complete.

$\square$

D.2 Proof of bounds of $\mathcal{S}(\mathbf{X}:Y)$

We show that,

$$\text{WMS}\,(\mathbf{X}:Y) \le \mathcal{S}\,(\mathbf{X}:Y) \le \mathcal{S}_{\max}\,(\mathbf{X}:Y) \le \text{I}(X_{1\ldots n}:Y) \ . \tag{48}$$

D.2.1 Proof that $\mathcal{S}_{\max}\,(\mathbf{X}:Y) \le \text{I}(X_{1\ldots n}:Y)$

Proof.

$$
\begin{aligned}
\mathcal{S}_{\max}\,(\mathbf{X}:Y) &\equiv \text{I}(X_{1\ldots n}:Y) - \sum_{y \in Y} \Pr(y) \max_i \text{I}(X_i:Y=y) && (49)\\
&= \text{I}(X_{1\ldots n}:Y) - \sum_{y \in Y} \Pr(y) \max_i \underbrace{\text{D}_{\text{KL}}\Big[\Pr\big(X_i|y\big)\,\big\|\,\Pr(X_i)\Big]}_{\ge 0} && (50)\\
&\phantom{= \text{I}(X_{1\ldots n}:Y) - \sum_{y \in Y} \Pr(y) \max_i \underbrace{}_{\ge 0}}\\
&\le \text{I}(X_{1\ldots n}:Y) \ . && (51)
\end{aligned}
$$

$\square$

D.2.2 Proof that $\mathcal{S}(\mathbf{X}:Y) \le \mathcal{S}_{\max}\,(\mathbf{X}:Y)$

We invoke the standard definitions of $\mathcal{S}$ and $\mathcal{S}_{\max}$,

$$
\begin{aligned}
\mathcal{S}(\mathbf{X}:Y) &\equiv \text{I}(X_{1\ldots n}:Y) - \text{I}_{\cup}(\mathbf{X}:Y) && (52)\\
\mathcal{S}_{\max}(\mathbf{X}:Y) &\equiv \text{I}(X_{1\ldots n}:Y) - \text{I}_{\max}(\mathbf{X}:Y) \ , && (53)
\end{aligned}
$$

where $\text{I}_{\cup}$ and $\text{I}_{\max}$ are defined as,

$$
\begin{aligned}
\text{I}_{\cup}\,(\mathbf{X}:Y) &= \mathbb{E}_Y\,\text{I}_{\cup}(\mathbf{X}:Y=y) && (54)\\
&= \mathbb{E}_Y \min_{\substack{X_{1\ldots n} \to Y \to y' \\ I(X_i:Y'=y')=I(X_i:Y=y) \quad \forall i}} \text{I}\big(X_{1\ldots n}:Y'=y'\big) && (55)\\
\text{I}_{\max}\,(\mathbf{X}:Y) &\equiv \mathbb{E}_Y \max_i \text{I}(X_i:Y=y) \ . && (56)
\end{aligned}
$$

Now we prove $\mathcal{S}(\mathbf{X}:Y) \le \mathcal{S}_{\max}(\mathbf{X}:Y)$ by showing that $\text{I}_{\cup}(\mathbf{X}:Y) \ge \text{I}_{\max}(\mathbf{X}:Y)$.

Proof.

$$
\begin{aligned}
\mathbb{E}_Y\,\text{I}_{\cup}\,(\mathbf{X}:Y=y) &\ge \mathbb{E}_Y\,\text{I}_{\max}\,(\mathbf{X}:Y=y) && (57)\\
\mathbb{E}_Y\,\big[\text{I}_{\cup}\,(\mathbf{X}:Y=y) - \text{I}_{\max}\,(\mathbf{X}:Y=y)\big] &\ge 0 \ . && (58)
\end{aligned}
$$

Now expanding $\text{I}_{\cup}(\mathbf{X}:Y=y)$ and $\text{I}_{\max}(\mathbf{X}:Y=y)$,

$$\mathbb{E}_Y\left[\left(\min_{\substack{X_{1\ldots n} \to Y \to y' \\ I(X_i:Y'=y')=I(X_i:Y=y) \quad \forall i}} \text{I}\big(X_{1\ldots n}:Y'=y'\big)\right) - \max_i \text{I}(X_i:Y=y)\right] \ge 0 \ . \tag{59}$$

We define the index $m \in \{1,\ldots,n\}$ such that $m = \operatorname{argmax}_i \text{I}(X_i:Y=y)$. The predictor with the most information about state $Y=y$ is thus X_m. This yields,

$$\mathbb{E}_Y\left[\left(\min_{\substack{X_{1\ldots n} \to Y \to y' \\ I(X_i:Y'=y')=I(X_i:Y=y) \quad \forall i}} \text{I}\big(X_{1\ldots n}:Y'=y'\big)\right) - \text{I}(X_m:Y=y)\right] \ge 0 \ . \tag{60}$$

23

The constraint $\mathrm{I}\big(X_i\!:\!Y'=y'\big) = \mathrm{I}(X_i\!:\!Y=y)$ entails that $\mathrm{I}(X_m\!:\!Y=y) = \mathrm{I}\big(X_m\!:\!Y'=y'\big)$. Therefore we can pull $\mathrm{I}(X_m\!:\!Y=y)$ inside the minimization as a constant,

$$\mathbb{E}_Y\left[\min_{\substack{X_{1\ldots n}\to Y\to y' \\ I(X_i:Y'=y')=I(X_i:Y=y) \quad \forall i}} \mathrm{I}\big(X_{1\ldots n}\!:\!Y'=y'\big) - \mathrm{I}\big(X_m\!:\!Y'=y\prime\big)\right] \geq 0\,. \qquad (61)$$

As X_m is a subset of predictors $X_{1\ldots n}$, we can substract it yielding,

$$\mathbb{E}_Y\left[\min_{\substack{X_{1\ldots n}\to Y\to y' \\ I(X_i:Y'=y')=I(X_i:Y=y) \quad \forall i}} \mathrm{I}\Big(X_{1\ldots n\backslash m}\!:\!Y'=y'\Big|X_m\Big)\right] \geq 0\,. \qquad (62)$$

The state-dependent conditional mutual information $\mathrm{I}\Big(X_{1\ldots n\backslash m}\!:\!Y'=y'\Big|X_m\Big)$ is a Kullback-Liebler divergence. As such it is nonnegative. Likewise the minimum of a nonnegative quantity is also nonnegative.

$$\mathbb{E}_Y\left[\underbrace{\min_{\substack{X_{1\ldots n}\to Y\to y' \\ I(X_i:Y'=y')=I(X_i:Y=y) \quad \forall i}} \underbrace{\mathrm{I}\Big(X_{1\ldots n\backslash m}\!:\!Y'=y'\Big|X_m\Big)}_{\geq 0}}_{\geq 0}\right] \geq 0\,. \qquad (63)$$

Finally, the expected value of a list of nonnegative quantities is itself nonnegative. And the proof that $\mathcal{S}(\mathbf{X}:Y) \leq \mathcal{S}_{\max}(\mathbf{X}:Y)$ is complete.

$\square$

D.2.3   Proof that $\mathrm{WMS}(\mathbf{X}:Y) \leq \mathcal{S}(\mathbf{X}:Y)$

We invoke the standard definitions of WMS and $\mathcal{S}$,

$$
\mathrm{WMS}(\mathbf{X}:Y) \;\equiv\; \mathrm{I}(X_{1\ldots n}:Y) - \sum_{i=1}^{n} \mathrm{I}(X_i:Y) \tag{64}
$$

$$
\mathcal{S}(\mathbf{X}:Y) \;\equiv\; \mathrm{I}(X_{1\ldots n}:Y) - \mathrm{I}_\cup(\mathcal{S}:Y) \tag{65}
$$

$$
= \; \mathrm{I}(X_{1\ldots n}:Y) - \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) \;. \tag{66}
$$

We prove the conjecture $\mathrm{WMS}(\mathbf{X}:Y) \leq \mathcal{S}(\mathbf{X}:Y)$ by showing,

$$
\min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) \leq \sum_{i=1}^{n} \mathrm{I}(X_i:Y) \;. \tag{67}
$$

Proof. Given:

$$
\min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_1:Y')=I(X_1:Y) \\ \vdots \\ I(X_n:Y')=I(X_n:Y)}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) \;, \tag{68}
$$

the individual constraint $\mathrm{I}\!\left(X_1:Y'\right) = \mathrm{I}(X_1:Y)$ can add at most $\mathrm{I}(X_1:Y)$ to $\mathrm{I}\!\left(X_{1\ldots n}:Y'\right)$. Therefore we can upperbound eq. (68) by dropping the constraint $\mathrm{I}\!\left(X_1:Y'\right) = \mathrm{I}(X_1:Y)$ and adding $\mathrm{I}(X_1:Y)$. This yields,

$$
\min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_1:Y')=I(X_1:Y) \\ \vdots \\ I(X_n:Y')=I(X_n:Y)}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) \leq \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_2:Y')=I(X_2:Y) \\ \vdots \\ I(X_n:Y')=I(X_n:Y)}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) + \mathrm{I}(X_1:Y) \;. \tag{69}
$$

Likewise, the righthand-side of eq. (69) can be upperbounded by dropping the constraint $\mathrm{I}\!\left(X_2:Y'\right) = \mathrm{I}(X_2:Y)$ and adding $\mathrm{I}(X_2:Y)$. This yields,

$$
\min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_2:Y')=I(X_2:Y) \\ \vdots \\ I(X_n:Y')=I(X_n:Y)}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) \leq \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_3:Y')=I(X_3:Y) \\ \vdots \\ I(X_n:Y')=I(X_n:Y)}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) + \mathrm{I}(X_1:Y) + \mathrm{I}(X_2:Y) \;. \tag{70}
$$

Repeating this process n times yields,

$$
\min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i}} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right) \;\leq\; \underbrace{\min_{X_{1\ldots n}\to Y\to Y'} \mathrm{I}\!\left(X_{1\ldots n}:Y'\right)}_{=0} + \sum_{i=1}^{n} \mathrm{I}(X_i:Y) \tag{71}
$$

$$
= \; \sum_{i=1}^{n} \mathrm{I}(X_i:Y) \;. \tag{72}
$$

And the proof is complete. $\qquad\square$

D.3 Proof that the union information is idempotent

We show that the synergistic mutual information $\mathcal{S}(\mathbf{X} : Y)$ is invariant under adding an additional predictor $X_i \in \{X_1, \ldots, X_n\}$. We assume without loss of generalization that the duplicated predictor is X_1. We will show that,

$$\mathcal{S}\left(\{X_1, \ldots, X_n, X_1\} : Y\right) = \mathcal{S}\left(\{X_1, \ldots, X_n\} : Y\right) . \tag{73}$$

Proof. We start with the expression for $\mathcal{S}\left(\{X_1, \ldots, X_n, X_1\} : Y\right)$,

$$\mathcal{S}\left(\{X_1, \ldots, X_n, X_1\} : Y\right) = \mathrm{I}(X_{1\ldots n}X_1{:}Y) - \min_{\substack{X_{1\ldots n}X_1 \to Y \to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i \\ I(X_1:Y')=I(X_1:Y)}} \mathrm{I}\left(X_{1\ldots n}X_1{:}Y'\right) . \tag{74}$$

The two mutual information terms do not change when duplicating predictor X_1. This yields,

$$\mathcal{S}\left(\{X_1, \ldots, X_n, X_1\} : Y\right) = \mathrm{I}(X_{1\ldots n}{:}Y) - \min_{\substack{X_{1\ldots n}X_1 \to Y \to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i \\ I(X_1:Y')=I(X_1:Y)}} \mathrm{I}\left(X_{1\ldots n}{:}Y'\right) . \tag{75}$$

Having the constraint $\mathrm{I}\left(X_1{:}Y'\right) = \mathrm{I}(X_1{:}Y)$ twice is superfluous. Therefore we can remove the latter one yielding,

$$= \mathrm{I}(X_{1\ldots n}{:}Y) - \min_{\substack{X_{1\ldots n}X_1 \to Y \to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i}} \mathrm{I}\left(X_{1\ldots n}{:}Y'\right) . \tag{76}$$

Finally, the Markov condition $X_{1\ldots n}X_1 \to Y \to Y'$ means that,

$$
\begin{aligned}
\Pr\left(x_{1\ldots n}x_1, y, y'\right) &= \Pr(x_{1\ldots n}, x_1)\Pr\left(y|x_{1\ldots n}, x_1\right)\Pr\left(y'|y\right) & (77)\\
&= \Pr(x_{1\ldots n}, x_1)\Pr\left(y|x_{1\ldots n}\right)\Pr\left(y'|y\right) & (78)\\
&= \Pr(x_{1\ldots n})\underbrace{\Pr\left(x_1|x_{1\ldots n}\right)}_{=1}\Pr\left(y|x_{1\ldots n}\right)\Pr\left(y'|y\right) & (79)\\
&= \Pr(x_{1\ldots n})\Pr\left(y|x_{1\ldots n}\right)\Pr\left(y'|y\right) ; & (80)
\end{aligned}
$$

which equates to the Markov condition $X_{1\ldots n} \to Y \to Y'$. Altogether, we end up with,

$$
\begin{aligned}
\mathcal{S}\left(\{X_1, \ldots, X_n, X_1\} : Y\right) &= \mathrm{I}(X_{1\ldots n}{:}Y) - \min_{\substack{X_{1\ldots n} \to Y \to Y' \\ I(X_i:Y')=I(X_i:Y) \quad \forall i}} \mathrm{I}\left(X_{1\ldots n}{:}Y'\right) & (81)\\
&= \mathcal{S}\left(\{X_1, \ldots, X_n\} : Y\right) , & (82)
\end{aligned}
$$

and the proof of eq. (73) is complete. $\qquad\square$

E   Nonessential proofs

These proofs demonstrate some interesting properties of synergistic mutual information, however, our paper does not rely on them.

E.1   Proof of zero synergy when $Y = X_{1\ldots n}$

Objective: Prove that,

$$\mathcal{S}\left(\{X_1,\ldots,X_n\}:Y\right) = 0 \quad \text{when } Y = X_{1\ldots n} .$$

Proof.

$$
\begin{aligned}
\mathcal{S}\left(\{X_1,\ldots,X_n\}:Y\right) &\equiv \mathrm{I}(X_{1\ldots n}:Y) - \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \ \ \forall i}} \mathrm{I}\left(X_{1\ldots n}:Y'\right) && (83) \\
&= \mathrm{I}(X_{1\ldots n}:X_{1\ldots n}) \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \ \ \forall i}} \mathrm{H}(X_{1\ldots n}) - \mathrm{H}\left(X_{1\ldots n}\big|Y'\right) \\
&= \mathrm{H}(X_{1\ldots n}) - \mathrm{H}(X_{1\ldots n}) + \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \ \ \forall i}} \mathrm{H}\left(X_{1\ldots n}\big|Y'\right) && (84) \\
&= \min_{\substack{X_{1\ldots n}\to Y\to Y' \\ I(X_i:Y')=I(X_i:Y) \ \ \forall i}} \mathrm{H}\left(X_{1\ldots n}\big|Y'\right) && (85)
\end{aligned}
$$

Setting $Y' = Y = X_{1\ldots n}$ puts $\mathrm{H}\left(X_{1\ldots n}\big|Y'\right) = 0$ and satisfies all constraints.

$$
\begin{aligned}
&= \ 0 . && (86)
\end{aligned}
$$

$\square$